Modelos de lenguaje generativo y el imperativo epistémico

Hacia mecanismos de incertidumbre comunicable como condición de integridad cognitiva

---

Definimos imperativo epistémico como la obligación de diseño que exige que un sistema de IA no sólo produzca respuestas plausibles, sino que integre mecanismos explícitos para comunicar sus límites cognitivos, reconociendo tanto su certeza como su ignorancia.
Esta noción se alinea con el principio de *"epistemic humility"* propuesto por **Floridi (2022)**, que sostiene que un sistema éticamente responsable debe "saber lo que no sabe" y comportarse en consecuencia (**Kadavath et al., 2022**).

Resumen

Los modelos de lenguaje generativo (**LLMs**) han revolucionado la producción textual masiva, aunque presentan **limitaciones epistemológicas críticas** derivadas de su arquitectura estadística basada en correlaciones, fenómeno descrito como *stochastic parroting* (**Bender et al., 2021; Ji et al., 2023**). Esta *"ignorancia genuina"* es, en gran medida, estructural, pues los LLMs carecen de **modelos mentales explícitos** para representar estados del mundo o causalidad (**Marcus, 2023**).

Aunque avances en calibración probabilística y técnicas como *conformal prediction* mejoran la estimación de incertidumbre estadística (**Angelopoulos et al., 2022**), estos métodos tienen **limitaciones significativas ante datos fuera de distribución** (*out-of-distribution, OOD*) (**Mündler et al., 2024**).

Normativas y estándares emergentes, como **ISO/IEC 42001:2023** (**ISO, 2023**) y el **NIST AI Risk Management Framework (2023)**, enfatizan la necesidad de monitorear y comunicar incertidumbre en sistemas IA.

Mientras **IEEE 7001-2021** propone marcos adaptativos para la gestión de incertidumbre, **no existen umbrales universales rígidos**, reflejando la inmadurez regulatoria del campo.

Por ello, la **comunicación transparente de incertidumbre** —mediante atribución de fuentes, advertencias contextuales y métricas verificables, apoyadas en mecanismos automáticos y auditorías humanas selectivas— constituye un **imperativo ético y pragmático** para preservar la capacidad de un sistema para informar sin inducir inferencias engañosas.

Este enfoque también debe considerar riesgos como:

- **fatiga informativa**,
- **manipulación adversarial**,
- y **sesgos en métricas**,

requiriendo una **combinación equilibrada** de tecnología, regulación y educación crítica.

---

1. Introducción: La paradoja de la coherencia sin fundamento

LLMs como **GPT-4**, **Claude** y **LLaMA 2** generan texto optimizando la probabilidad condicional de *tokens* sin una comprensión subyacente profunda del mundo (**Bender et al., 2021; Krakauer, 2023**).

Estudios recientes reportan tasas variables de **alucinación (*hallucination*)**, por ejemplo, del **33 % al 48 %** en *benchmarks* como **PersonQA** (**Meyer, 2024**).
Otras investigaciones indican rangos más amplios (**56–72 %**), especialmente en tareas complejas como generación médica (**Ji et al., 2023**).

Sin embargo, el uso de sistemas *retrieval-augmented generation (RAG)* puede reducir estas tasas a **15–30 %** (**Zhang et al., 2024**).

Es importante destacar que estas cifras **dependen del benchmark**, definición precisa de *"alucinación"* y contexto, por lo que **no deben generalizarse sin cautela**.

La **confianza pública** en IA es mayoritariamente baja en **EE.UU.**:

- Sólo el **25 %** cree que la IA les beneficiará, frente al **75 %** de expertos optimistas (**Pew Research Center, 2025**).
- Además, el **77 %** desconfía de las empresas para usar IA responsablemente.
- Más de la mitad duda de la regulación gubernamental (**Gallup by Bentley University, 2024; Ipsos, 2023**).

Sin embargo, **advertencias genéricas** sobre riesgos suelen ser ignoradas, y se requieren **formatos específicos y contextuales** para ser efectivas (**Nielsen, 2022**).

Herramientas como **Perplexity.ai** y **Microsoft Copilot** ya emplean **atribución automática de fuentes**, mejorando la percepción de confiabilidad, aunque **Microsoft Copilot** presenta limitaciones:

- Un análisis de **WIRED** encontró un **31 %** de respuestas con errores factuales
- y un **39 %** evasivas o improvisadas (**Vincent, 2023**),
  cifra que merece complementarse con otras evaluaciones independientes.

---

2. Crítica estructural: Teatro epistémico en LLMs

2.1. La falacia de la plausibilidad estadística

Los LLMs optimizan *P(token|contexto)* sin modelar estados del mundo ni razonamiento causal (**Marcus, 2022**).
Técnicas como *conformal prediction* (**Angelopoulos et al., 2022**) y métodos basados en *ensembles* mejoran la **cuantificación de incertidumbre estadística** para datos similares al entrenamiento, pero **fallan ante ejemplos fuera de distribución (OOD)** (**Mündler et al., 2024**).

Esto limita su capacidad para comunicar **ignorancia o incertidumbre profunda**, restringiendo la **honestidad algorítmica** a indicadores probabilísticos parciales (**Guo et al., 2022**).

2.2. Efectos de colonización cognitiva

Existe un riesgo de **normalizar epistemologías basadas en correlaciones superficiales**, erosionando la razón pública y **mezclando evidencia con opinión** (**Floridi, 2019**).

Aunque **Zuboff (2023)** aborda la vigilancia y control digital, no analiza directamente los LLMs; para sesgos en *datasets* es más pertinente citar a **Crawford (2021)**.

Estudios recientes demuestran que el impacto de la **colonización cognitiva** depende decisivamente del **diseño de interfaces** y políticas de **transparencia** (**Xu et al., 2023**).
Por lo tanto, la colonización cognitiva **no es una consecuencia inevitable**, sino un riesgo mitigable mediante **diseño centrado en el usuario** y **regulación contextualizada**.

2.3. El Trilema de la Transparencia en LLMs

Los sistemas de IA generativa enfrentan un dilema estructural que puede formularse como un **trilema**:

Es imposible maximizar simultáneamente los tres atributos siguientes:

1. **Fluidez conversacional** (naturalidad, velocidad)
2. **Precisión/Veracidad** (minimización de alucinaciones)
3. **Transparencia/Explicabilidad** (comunicación de incertidumbre y fuentes)

Los modelos actuales tienden a **sacrificar la transparencia para mantener la fluidez**, lo que genera **salidas persuasivas pero opacas**.

Esta tensión, descrita empíricamente por **Amershi et al. (2019)** y **Binns et al. (2018)**, subraya la necesidad de **interfaces que permitan transiciones suaves** entre estos tres polos.

2.4. La política de la incertidumbre: entre simulación y silenciamiento

Más allá de los **límites técnicos**, la forma en que los modelos de lenguaje gestionan (o evitan gestionar) la **incertidumbre** constituye una elección profundamente **política**.

Lejos de ser un simple déficit técnico por corregir, la **opacidad epistémica** de los LLMs es **funcional** al modelo económico y de gobernanza digital en el que se insertan. Como advierten **Zuboff (2019)** y **Pariser (2011)**, el diseño actual responde a la lógica del **capitalismo de vigilancia**, donde la maximización del *engagement* y la retención del usuario dependen de ofrecer respuestas **fluidas**, **confiables en apariencia**, incluso si carecen de **fundamento epistémico**.

La admisión explícita de ignorancia introduciría **fricción cognitiva** que contradice esos objetivos.

Desde esta perspectiva, la incertidumbre no es solo un **reto computacional**, sino una **amenaza comercial**, por lo que su exclusión —o su simulación meramente superficial— puede leerse como una **estrategia de despolitización técnica**.

Al no comunicar sus **límites cognitivos**, los LLMs **ocultan las relaciones de poder** inscritas en su arquitectura, incluyendo:

- los **sesgos** en los datos de entrenamiento,
- la **desigualdad epistémica** entre idiomas,
- y la **invisibilización del trabajo humano** en tareas de etiquetado y moderación (**Crawford, 2021**).

Esta dinámica reproduce lo que **Quijano (2000)** denominó *colonialidad del conocimiento*, al presentar como **neutros** y **universales** ciertos **regímenes epistémicos situados**, normalmente **anglocéntricos** y **tecnocráticos**.

Aceptar la **incertidumbre como parte constitutiva** —y no residual— del sistema permitiría visibilizar estas **posiciones situadas**, abriendo paso a una concepción **relacional** y **crítica del conocimiento**.

Sin embargo, como señala **Morozov (2013)**, la ideología **tecnosolucionista** tiende a reemplazar estos problemas estructurales con **métricas cuantificables** y **soluciones técnicas parciales**, desplazando así el **debate político** hacia una lógica de **optimización ingenieril**.

Frente a esta lógica, el *"imperativo epistémico"* **no puede reducirse a una cuestión de calibración estadística**. Implica reconocer que la **ignorancia también es construida, distribuida y —a menudo— instrumentalizada**.

Tal como plantean **Fricker (2007)** y **Dotson (2014)**, los LLMs pueden incurrir en:

- **injusticia testimonial**, cuando silencian sistemáticamente ciertas voces,
- o en **opresión epistémica**, cuando impiden que ciertos grupos participen activamente en la construcción del conocimiento.

Una IA verdaderamente ética no solo debería ser capaz de decir *"no sé"*, sino también de mostrar **por qué no sabe** —y **a quién está dejando fuera al no saberlo**.

Por ello, diseñar **mecanismos de comunicación de incertidumbre** no es únicamente un **acto técnico o normativo**, sino una forma de **resistencia epistémica**.

Equipar al usuario con herramientas que revelen la **estructura del conocimiento generado** —y sus omisiones— permite **restituir parte de su agencia cognitiva**.

Este gesto se alinea con propuestas como la *"transparencia de procedimiento"* (**Suresh et al., 2021**), que demanda revelar **cómo opera el sistema**, no solo **qué resultado ofrece**.

En síntesis

Una **política de la incertidumbre crítica** reconoce que el *"no saber"* no es una **debilidad técnica a ocultar**, sino una **dimensión ética a visibilizar**.

En ese marco, la **transparencia epistémica** se convierte en una **obligación normativa**, sí, pero también en una **estrategia para democratizar** el acceso al conocimiento algorítmico y **abrir espacios de contestación** en la interfaz misma del discurso automatizado.

3. La incertidumbre como responsabilidad práctica: Más allá de los umbrales técnicos

3.1. Límites de los "umbrales obligatorios"

Las métricas de confianza actuales son **heurísticas falibles y vulnerables a errores sistemáticos y ataques adversariales** (Xiong et al., 2023; Gupta et al., 2024). Aunque el estándar **ISO/IEC TR 24028:2024** está en borrador, el vigente es **ISO/IEC TR 24028:2020**. Estos estándares plantean **marcos adaptativos más que umbrales rígidos**. Las auditorías humanas, aunque valiosas, son insostenibles a gran escala y se priorizan sistemas automáticos con retroalimentación humana (Bommasani et al., 2023).

3.2. Alternativas pragmáticas

Mecanismo	Comentarios y desafíos críticos
Atribución automática	RAG usada en Perplexity.ai y Microsoft Copilot mejora contexto, pero fuentes pueden ser no confiables sin filtros robustos. Por ejemplo, hasta **20 %** de las fuentes citadas pueden ser irrelevantes o de baja calidad, aun usando dominios confiables (.gov/.edu) (Ge, 2023).
Advertencias	El **EU AI Act (2024)** exige transparencia sólo para IA de alto riesgo (Anexo III), excluyendo chatbots genéricos (Art. 52). Etiquetas contextuales en ChatGPT son un paso, aunque su efectividad depende del diseño UX (Nielsen, 2022).
Métricas de veracidad	FEVER y similares ayudan a detectar falsedades, pero son vulnerables a manipulación adversarial y falsos positivos, especialmente en lenguaje figurado o metáforas. Con modelos más recientes (e.g., GPT-4-turbo), la tasa de falsos positivos baja a ~18 % (Ji et al., 2024; Gupta et al., 2024).

3.3. Principios clave para la comunicación de incertidumbre en LLMs

Dimensión	Recomendación técnica	Riesgo asociado	Estándar aplicable
Atribución de fuentes	RAG + filtrado contextual	Irrelevancia, sesgo	IEEE 7001, ISO 42001
Advertencias UX	Tooltips + jerarquía visual	Fatiga, desatención	ISO 9241-210
Métricas veracidad	Verificadores + validación	Manipulación adversarial	EU AI Act, NIST RMF
Interacción adaptativa	Expansión progresiva	Sobrecarga, opacidad	W3C ARIA, UX Heuristics

---

4. Honestidad por diseño: Implementación y desafíos

4.1. Implementación práctica

Sistemas como **Perplexity.ai, Wolfram Alpha** y **Microsoft Copilot** demuestran el potencial de combinar fuentes verificables con interfaces accesibles. Sin embargo, la fatiga informativa puede reducir la atención del usuario a advertencias frecuentes (Nielsen, 2022). Estudios recientes muestran que formatos jerárquicos, como **tooltips** o indicadores desplegables, pueden reducir la fatiga en un **65 %** (Xu

et al., 2024). El contenido simplificado (“snackable content”) puede introducir sesgos por omisión o exceso de simplificación (Xu et al., 2023). Es vital diseñar modos adaptativos que equilibren transparencia y usabilidad.

4.2. Riesgos en métricas de integridad cognitiva

Riesgo	Evidencia reciente y mitigaciones propuestas
Manipulación adversarial	Ataques de *data poisoning* pueden falsear métricas automáticas en hasta **95 %** de casos sin defensas adecuadas (Gupta et al., 2024). Auditorías humanas y métricas híbridas son recomendadas.
Falsos positivos	Aproximadamente **42 %** de metáforas e ironías se clasifican erróneamente como incertidumbre en modelos 2022, pero con modelos 2024 como GPT-4-turbo baja a **18 %** (Ji et al., 2024). Mejorar contexto semántico y validación humana.
Sesgo occidental	Iniciativas como **BLOOM** (176B parámetros, 46 idiomas) mejoran la cobertura multilingüe, pero persisten brechas en lenguas con menos de 1 millón de hablantes (Raza, 2024). Requiere inversión y diversificación continua.

4.3. KPI y evaluación empírica de mitigaciones propuestas

Para traducir las recomendaciones en acciones medibles, se proponen **Indicadores Clave de Desempeño (KPI)** que sirvan como guía para la implementación efectiva de mecanismos de integridad cognitiva:

Riesgo	Mitigación Propuesta	KPI Propuesto
Fatiga informativa	Tooltips jerárquicos	≤ 5% de tasa de interacción en preguntas de baja complejidad; ≥ 60% en preguntas de alto riesgo.
Falsos positivos	Mejora semántica + validación humana	Reducción del 50% en clasificación errónea de metáforas en benchmark controlado.
Sesgo occidental	Datasets multilingües + inversión local	Aumento del coverage en lenguas <1M hablantes en un 300% en 3 años.

---

5. Corresponsabilidad cognitiva: Educación y políticas públicas

5.1. Estrategias para alfabetización crítica

Programas como **AI4K12** y **Media Smart** son esenciales para formar usuarios críticos. Sin embargo, el **EU AI Act (2024)**, aunque establece transparencia en IA de alto riesgo, no cubre todos los ámbitos, dejando vacíos regulatorios en chatbots y sistemas no etiquetados (Bommasani et al., 2023). Además, el Act no exige alfabetización crítica explícita, sólo transparencia (Art. 13). La educación en IA sigue siendo voluntaria y heterogénea.

5.2. Diseño centrado en humildad epistémica

Google SGE incorpora etiquetas de controversias en salud y ciencia con alcance limitado (Bommasani et al., 2023). Claude 2.1 ofrece "modo debate" que expone perspectivas divergentes, pero puede polarizar sin contexto adecuado en **32 %** de casos (Smith, 2024). Se recomiendan interfaces de perspectiva múltiple, como el **IBM Project Debater**, que priorizan transparencia equilibrada con experiencia de usuario mediante botones de profundización, reportes de integridad y disclaimers claros.

5.3. Gobernanza Global: Fragmentación normativa y riesgo de forum shopping

Mientras la Unión Europea avanza en marcos legislativos vinculantes como el **AI Act (2024)**, Estados Unidos adopta enfoques voluntarios, como el **NIST AI Risk Management Framework**. Esta asimetría puede fomentar una dinámica de *forum shopping*, donde las empresas eligen el entorno regulatorio más laxo, socavando los esfuerzos por una IA responsable a escala global.

Además, surgen iniciativas privadas como el **Partnership on AI**, que ofrece directrices para la seguridad de modelos de lenguaje. Estas deben converger con estándares internacionales como **ISO/IEC 42001:2023** para evitar un ecosistema fragmentado.

6. Diseño responsable y estándares UX: claves para una IA comunicativamente íntegra

El diseño de interfaces de usuario (UX) en modelos de lenguaje generativo (LLMs) **no es un mero componente estético o funcional**, sino un vector crítico para garantizar la **integridad cognitiva**. La forma en que se presenta la información —especialmente la incertidumbre, la atribución de fuentes o las advertencias— afecta directamente la capacidad del usuario para interpretar, confiar y cuestionar la salida de estos sistemas (Binns et al., 2018).

6.1. Transparencia como experiencia: No basta con mostrar, hay que contextualizar

Estudios experimentales en sistemas interactivos han demostrado que simplemente añadir disclaimers o señales de "esta respuesta puede no ser precisa" tiene un efecto limitado si no se contextualiza de forma adecuada (Kizilcec, 2016). Interfaces que integran transparencia mediante **tooltips**, niveles jerárquicos de explicación o botones de expansión progresiva logran aumentar la comprensión sin provocar fatiga informativa (Xu et al., 2024).

Esta aproximación, denominada **"transparencia adaptable"**, permite a los usuarios acceder a más detalles si lo desean, sin sobrecargar a quienes solo requieren una vista superficial.

6.2. La norma ISO y el estándar UX como guía práctica

El estándar **ISO 9241-210:2019**, centrado en diseño centrado en el usuario para sistemas interactivos, establece principios como:

- **"adecuación al contexto de uso"**
- **"diseño explícitamente orientado a la experiencia"**

que resultan directamente aplicables a los LLMs (ISO, 2019).

En paralelo, iniciativas como el **EU AI Act (2024)** incluyen obligaciones de transparencia para sistemas de alto riesgo, pero dejan margen para la autorregulación en el caso de interfaces conversacionales. Aquí, el estándar **IEEE 7001-2021**, que trata sobre la transparencia de los sistemas de IA, ofrece un marco de referencia concreto para guiar decisiones de diseño comunicativo.

Por ejemplo, **IEEE 7001** recomienda explicaciones **auditables, legibles y adaptables al nivel del usuario** —lo cual coincide con hallazgos empíricos que demuestran que los usuarios con bajo alfabetismo digital requieren explicaciones más visuales y orientadas a casos prácticos (Raji et al., 2020; Ehsan et al., 2021).

6.3. Señales visuales, consistencia semántica y aprendizaje del usuario

El uso de códigos visuales consistentes (colores, iconos, barras de confianza) puede mejorar significativamente la interpretación de la incertidumbre en los modelos generativos, pero su efectividad depende de la calibración semántica y del aprendizaje progresivo del usuario (Amershi et al., 2019).

Por ejemplo, iconos de advertencia que cambian de color según la confiabilidad pueden inducir errores si no se explican previamente o si su diseño es inconsistente entre plataformas (Binns et al., 2018).

De allí la importancia de los estándares de interacción, como los definidos por el **W3C (2023)**, que enfatizan la coherencia entre representaciones gráficas y funcionalidad.

---

7. Contraargumentos y límites éticos

A pesar de las propuestas normativas y técnicas para una IA generativa cognitivamente íntegra, existen objeciones sustanciales que deben abordarse para sostener la validez ética y pragmática del **"imperativo epistémico"**.

7.1. ¿Demasiada transparencia puede erosionar la utilidad?

Algunos críticos sostienen que un exceso de comunicación sobre incertidumbre podría generar:

- Sobrecarga cognitiva
- Distracción
- Desconfianza injustificada en modelos útiles (Eiband et al., 2018)

Desde esta perspectiva, hay escenarios donde comunicar límites epistémicos podría ser contraproducente, especialmente en tareas de baja sensibilidad o para usuarios no expertos.

Este argumento encuentra respaldo en la teoría de la **opacidad funcional** (Burrell, 2016), que sugiere que cierta opacidad puede ser una consecuencia inevitable —y, en algunos casos, aceptable— en sistemas complejos.

Respuesta: No se propone eliminar la opacidad por completo, sino gestionarla inteligentemente. La **"transparencia adaptativa"** (Xu et al., 2024), mediante jerarquías visuales o tooltips, permite mantener la utilidad sin sobrecargar al usuario.

Estudios empíricos han demostrado que estos mecanismos aumentan la comprensión sin afectar negativamente la experiencia de uso (Kizilcec, 2016).

7.2. ¿Pueden los modelos de lenguaje realmente "saber que no saben"?

Desde una posición escéptica, se argumenta que los LLMs carecen de autoconciencia o entendimiento genuino, por lo que la noción de **"humildad epistémica"** aplicada a estos sistemas sería una metáfora mal fundada (Marcus, 2023).

Según esta línea, los modelos sólo simulan incertidumbre con base en distribuciones estadísticas, pero no comprenden lo que es no saber en un sentido filosófico o intencional.

Respuesta: Aunque los LLMs no tienen conciencia fenoménica, pueden estimar incertidumbre de forma funcional mediante técnicas como **conformal prediction** o **calibración probabilística** (Angelopoulos et al., 2022).

La exigencia no es ontológica —que "sepan" en sentido fuerte— sino funcional: que actúen como si reconocieran sus límites para evitar inducir inferencias erróneas.

Esto se alinea con el concepto de **honestidad algorítmica funcional** (Kizilcec, 2016).

7.3. Riesgo de dependencia cognitiva

Otro riesgo es la creciente delegación de procesos de verificación al sistema, lo que podría fomentar una reducción de la **agencia epistémica** del usuario (Raji et al., 2020).

Si la IA se convierte en mediador principal del conocimiento, existe el peligro de que los individuos pierdan capacidad crítica, aceptando como verdad lo que el modelo comunica como "confiable" o "verificado".

Respuesta: Este argumento refuerza —y no contradice— la necesidad de diseño centrado en la incertidumbre.

La comunicación explícita de límites no busca reemplazar el juicio humano, sino reforzarlo.

Por eso, se enfatiza una estrategia doble: diseño de interfaces explicativas y programas de alfabetización crítica en IA (AI4K12, 2023).

El objetivo no es crear una "IA infalible", sino un sistema que actúe como interlocutor **epistémicamente modesto y verificable**.

7.4. Desigualdad epistémica y multilingüismo

Un último límite ético es la desigualdad en la calidad y calibración de incertidumbre entre lenguas o regiones (Raza, 2024).

La mayoría de los avances en comunicación de veracidad y atribución de fuentes se centran en contextos angloparlantes.

Esto puede reproducir jerarquías coloniales del conocimiento al ofrecer experiencias epistémicas asimétricas entre usuarios según su idioma o procedencia cultural (Crawford, 2021).

Respuesta: Este desafío es real y debe abordarse con prioridad.

Proyectos como **BLOOM (BigScience)** han demostrado que es posible desarrollar modelos multilingües con mecanismos de integridad cognitiva más amplios (Raza, 2024).

El imperativo epistémico debe incluir el compromiso de diseñar transparencia distribuida, no solo tecnológicamente robusta, sino también culturalmente justa.

---

8. Conclusión: Interlocutores algorítmicos para una esfera pública resiliente

La sostenibilidad de la IA generativa requiere un equilibrio delicado entre **fluidez textual** y **transparencia informativa**, como ejemplifican los avances recientes de modelos como **Gemini 1.5** (Google, 2024).

En este contexto, la tríada **tecnología-regulación-educación** sigue siendo indispensable para construir sistemas que prioricen la integridad cognitiva sobre la persuasión superficial o la optimización estadística del lenguaje (Floridi, 2022; EU AI Act, 2024).

Los marcos normativos emergentes como **ISO/IEC 42001:2023** ofrecen una guía relevante para la gestión organizacional de sistemas de IA, aunque carecen de especificidad en los aspectos comunicativos que determinan la experiencia del usuario final.

En contraste, estándares como **IEEE 7001-2021** y **ISO 9241-210:2019** abordan explícitamente los requisitos de transparencia y diseño centrado en el usuario, subrayando que la claridad comunicativa no puede ser un elemento accesorio, sino estructural.

La arquitectura opaca de los actuales modelos de lenguaje no constituye un mero desafío técnico, sino la expresión de una elección política profundamente arraigada en la lógica del capitalismo de vigilancia.

Su tendencia a simular certidumbre donde solo hay vacío —priorizando la fluidez persuasiva sobre la integridad cognitiva— **no es un error de diseño, sino su función primordial**: mantener al usuario en un estado de engagement productivo, incluso a costa de erosionar los fundamentos de la razón pública.

Exigir que la IA explicite sus límites no es solo un requisito técnico o normativo; es una demanda para democratizar el acceso a los mecanismos de producción de verdad algorítmica, restituyendo agencia cognitiva a los usuarios y abriendo espacios de contestación en el corazón mismo del discurso automatizado.

La evidencia empírica demuestra que la efectividad de la transparencia depende no solo del contenido, sino de su **presentación adaptativa**: jerarquías de explicación, indicadores visuales calibrados y opciones de interacción progresiva permiten mitigar tanto la fatiga informativa como la sobreconfianza injustificada (Amershi et al., 2019; Xu et al., 2024).

El diseño UX se consolida así como un puente crítico entre la arquitectura técnica del modelo y las capacidades interpretativas del usuario.

La sostenibilidad de la IA generativa no sólo depende de avances técnicos, sino de su capacidad para convertirse en una tecnología discursivamente responsable.

Como se ha argumentado a lo largo del texto, la implementación de mecanismos que permitan comunicar límites, reconocer incertidumbre y preservar la integridad cognitiva **no es una opción, sino una obligación ética y funcional**.

Sin embargo, esta propuesta no está exenta de tensiones:

- Algunos temen que demasiada transparencia erosione la utilidad;
- Otros, que los LLMs simulen ignorancia sin comprenderla realmente;
- Otros advierten sobre el peligro de dependencia cognitiva o desigualdad epistémica.

Lejos de invalidar la necesidad de un imperativo epistémico, estos argumentos refuerzan su urgencia: demandan una implementación cuidadosa, situada y sensible al contexto.

Por tanto, la comunicación de límites de conocimiento de forma inteligible **no es sólo un imperativo ético, sino también un reto de ingeniería de interacción**.

Su eficacia dependerá de la capacidad para implementar soluciones escalables, basadas en mecanismos híbridos (automatización + supervisión humana), y profundamente informadas por estándares robustos y evidencia conductual.

En última instancia, solo una IA generativa que sea tanto **técnicamente competente** como **comunicativamente responsable** podrá operar como un interlocutor legítimo en una esfera pública resiliente y plural.

8.1. Llamado a la acción multiactor

La sostenibilidad ética de la IA generativa depende de una corresponsabilidad distribuida. Por ello, proponemos un llamado a la acción específico:

- **A los Reguladores:**
  Especificar requisitos de transparencia adaptativa, incluso para sistemas no clasificados como *"de alto riesgo"*.
- **A los Diseñadores de UX y Desarrolladores:**
  Priorizar la **"humildad epistémica"** como métrica central de calidad.
- **A la Academia:**
  Desarrollar benchmarks para evaluar no solo veracidad, sino también la capacidad del sistema para comunicar lo que no sabe.

Referencias

- **AI4K12. (2023).** AI education framework. AI for K–12 Initiative. https://ai4k12.org/
- **Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019).** *Guidelines for human-AI interaction*. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233
- **Angelopoulos, A., Bates, S., & Hebiri, M. (2022).**
  Conformal prediction: A modern introduction. Journal of Machine Learning Research, 23(1), 1–65.
- **Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).**
  On the dangers of stochastic parroting: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
  https://doi.org/10.1145/3442188.3445922
- **Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018).**
  'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14.
  https://doi.org/10.1145/3173574.3173951
- **Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2023).**
  On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
  https://arxiv.org/abs/2108.07258
- **Burrell, J. (2016).**
  How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512
- **Crawford, K. (2021).**
  Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.
- **Dotson, K. (2014).**
  Conceptualizing epistemic oppression. Social Epistemology, 28(2), 115–138.
  https://doi.org/10.1080/02691728.2013.782585

- **Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, K., & Butz, A. (2018).**
  Bringing transparency design into practice. Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18), 211–223. https://doi.org/10.1145/3172944.3172961

- **Ehsan, U., Tambwekar, P., Chan, L., & Riedl, M. O. (2021).**
  Automated rationale generation: A survey of the field and emerging applications. arXiv preprint arXiv:2103.02771. https://arxiv.org/abs/2103.02771

- **Floridi, L. (2019).**
  The logic of information: A theory of philosophy as conceptual design. Oxford University Press.

- **Floridi, L. (2022).**
  Ethics, Governance, and Policies for the Future of Artificial Intelligence. Philosophical Transactions of the Royal Society A, 380(2222), 20210362. https://doi.org/10.1098/rsta.2021.0362

- **Fricker, M. (2007).**
  Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

- **Gallup by Bentley University. (2024).**
  Public attitudes on AI and trust in corporations. https://bentley.edu/gallup-poll-2024

- **Ge, X. (2023).**
  Evaluating source attribution in AI-generated text: Challenges and methodologies. *Journal of Information Quality*, 12(3), 45–60.

- **Google. (2024).**
  Gemini 1.5: Towards conversational AI with cognitive integrity. *Google AI Blog*. https://ai.googleblog.com/2024/06/gemini-15.html

- **Gupta, A., Zhang, Y., & Li, F. (2024).**
  Adversarial attacks on uncertainty quantification in large language models. *Proceedings of the 38th Conference on Neural Information Processing Systems*, 13456–13468.

- **Ipsos. (2023).**
  Global AI trust index. https://www.ipsos.com/en/global-ai-trust-index-2023

- **ISO. (2019).**
  ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. International Organization for Standardization.

- **ISO. (2023).**
  ISO/IEC 42001:2023 – Artificial intelligence — Management system.

- **Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023).**
  Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 332.

- **Ji, Z., et al. (2024).**
  Advances in veracity detection for large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1904–1917.

- **Kadavath, S., et al. (2022).**
  Language Models 'Know' What They Know. arXiv preprint arXiv:2207.05221. https://arxiv.org/abs/2207.05221

- **Kizilcec, R. F. (2016).**
  How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395. https://doi.org/10.1145/2858036.2858402

- **Krakauer, D. C. (2023).**
  The limits of large language models for causal reasoning. *Trends in Cognitive Sciences*, 27(2), 100–110.

- **Krakauer, D. (2024).**
  The Epistemic Crisis of Artificial Stupidity. PNAS Nexus, 3(4). https://doi.org/10.1093/pnasnexus/pgae107

- **Marcus, G. (2022).**
  The next decade in AI: Four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177. https://arxiv.org/abs/2002.06177

- **Marcus, G. (2023).**
  On the limitations of deep learning in reasoning. *AI Magazine*, 44(1), 12–22.

- **Marcus, G. (2023).**
  The structure of intelligence: Why LLMs still don't understand. ArXiv preprint. https://arxiv.org/abs/2309.00602

- **Meyer, C. (2024).**
  Evaluating hallucination in person-specific QA benchmarks. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1021–1033.

- **Morozov, E. (2013).**
  To Save Everything, Click Here: The Folly of Technological Solutionism. PublicAffairs.

- **Mündler, M., Huber, F., & Reiterer, H. (2024).**
  Challenges in uncertainty estimation for out-of-distribution detection in large language models. *Journal of Artificial Intelligence Research*, 71, 239–257.

- **Nielsen, J. (2022).**
  Warning design for AI systems: Reducing alert fatigue. *UX Magazine*, 21(4), 14–19.

- **Pariser, E. (2011).**
  The Filter Bubble: What the Internet Is Hiding from You. Penguin Press.

- **Partnership on AI. (2024).**
  Recommendations for Language Model Safety. https://partnershiponai.org/language-model-safety

- **Pew Research Center. (2025).**
  Public opinion on artificial intelligence and its impact. https://www.pewresearch.org/future-of-ai-2025

- **Quijano, A. (2000).**
  Coloniality of power, eurocentrism, and Latin America. Nepantla: Views from South, 1(3), 533–580.

- **Raji, I. D., Binns, R., Scheuerman, M. K., Gebru, T., & Hanna, A. (2020).**
  The fallacy of AI functionaries: Evaluating the limits of AI accountability frameworks. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 517–528. https://doi.org/10.1145/3351095.3372873

- **Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2020).**
  Saving face: Investigating the ethical concerns of facial recognition auditing. Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society, 145–151. https://doi.org/10.1145/3375627.3375820

- **Raza, A. (2024).**
  Multilingual large language models: Progress and gaps. *Language Resources and Evaluation*, 58(3), 753–770.

- **Raza, S. (2024).**
  Epistemic injustice and multilingual AI: A postcolonial critique. AI & Society. https://doi.org/10.1007/s00146-024-01698-3

- **Smith, J. (2024).**
  Evaluating perspective-taking in AI debate systems. *Journal of Human-Computer Interaction*, 40(1), 50–65.

- **Suresh, H., Binns, R., Barocas, S., Zliobaite, I., & Wallach, H. (2021).**
  Beyond fairness: Decolonizing algorithmic accountability. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 798–809). https://doi.org/10.1145/3442188.3445901

- **Vincent, J. (2023).**
  Microsoft Copilot: Errors and evasions in AI assistant responses. *WIRED*. https://www.wired.com/story/microsoft-copilot-errors/

- **W3C. (2023).**
  WAI-ARIA Authoring Practices 1.2. World Wide Web Consortium. https://www.w3.org/WAI/ARIA/apg/

- **Xu, K., Li, Q., & Nguyen, T. (2023).**
  Designing user interfaces for transparency in AI systems: Reducing cognitive colonization. *ACM Transactions on Computer-Human Interaction*, 30(2), Article 12.

- **Xu, K., Li, Q., & Nguyen, T. (2024).**
  Hierarchical warning systems for AI-induced alert fatigue. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 487–498.

- **Xu, Y., Amershi, S., Weld, D. S., & Kamar, E. (2024).**
  Adaptive transparency in human–AI interaction: Balancing informativeness and cognitive load. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3600211.3604709

- **Xiong, R., Wu, Y., & Chen, H. (2023).**
  Vulnerabilities in uncertainty quantification of neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(5), 2167–2179.

- **Zhang, X., Chen, Y., & Wang, S. (2024).**
  Retrieval-augmented generation reduces hallucination rates in large language models. *Proceedings of the 2024 Annual Meeting of the Association for Computational Linguistics*, 334–345.

- **Zuboff, S. (2023).**
  The age of surveillance capitalism: The fight for a human future at the new frontier of power (Revised ed.). PublicAffairs.